Case Study: Strathprints, Institutional Repository, University of Strathclyde

Primary contact for case study: Alan Slevin, Institutional Repository Co-ordinator

Additional input from: Alan Dawson, Strathprints Technical Developer

1. Overview

The general aim of the case study is to explore the real-world potential for implementing identified Greening Information Management (GIM) methods. Each case study undertaken will determine current information management practice across a specific information service/collection within Higher Education Institutions (HEIs). It will then assess the feasibility of implementing GIM methods within that environment and consider the costs and benefits to the organisation as a result of such implementation(s).

2. Introduction

The University of Strathclyde was founded by John Anderson, a Professor of Natural Philosophy at Glasgow University and a visionary who wanted to provide 'a place of useful learning', that would be accessible to all. Originally 'Anderson's University', Strathclyde opened in 1796. By the 1890s, it was well-established as a technological institution with a reputation for research and learning.

Strathprints[1], the University of Strathclyde's Institutional Repository, is an open access[2] repository, intended to provide access to all of the University of Strathclyde's research outputs and other material produced by University staff. A mandate was recently passed by senior officers within the University, requesting that all research output be deposited in Strathprints, either by self-archiving or via a departmental proxy.

The University of Strathclyde runs two independent repository platforms – EPrints and Digitool. Strathprints, the main focus of the case study, runs on EPrints[3] software. Internal policy dictates what type of resources go into each. For the purpose of the current case study, our main focus is the EPrints repository – Strathprints, although the Digitool repository will be referred to.

A third publications system was developed within the Engineering Department at the University to assist in the preparation of the institution's RAE 2008 submission. Due to the infancy of Strathprints at the time of RAE preparation, and the lack of a strategy to accommodate measures of esteem and environment within Strathprints alongside research outputs, the University RAE team opted to handle RAE-related elements via a customised database within the Engineering Department, referred to as the RM (research metrics) database.

---

[1] http://strathprints.strath.ac.uk/
[2] http://www.jisc.ac.uk/publications/documents/pub_openaccess_v2.aspx
[3] http://www.EPrints.org/

Currently, the RM database content is being transferred to Strathprints, and Strathprints will be made the primary repository for research output, with contributors being asked to submit their works to this repository only.

3. Standards

Set metadata fields are populated according to established conventions, particular to specific item types. Strathprints is OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) compliant, meaning it can be harvested by external services and aggregators (e.g. OAIster). LCC (Library of Congress Classification) is used as the basis of a subject browse menu.

4. Phase 1: Examining the current IE

Strathprints currently holds full-text versions (where possible and permissible) of all University staff research outputs. It also holds metadata records corresponding to each item. Where full-text versions are not available, or cannot be made available on an open access database, owing to legal or other restrictions, metadata-only records are held to provide a description of items. Item types include journal articles, conference papers, workshop items, book chapters, teaching resources, reviews and patents.

Authority files are currently being developed for author names. This involves the creation of a unique form to link existing variant name formats together, so that each individual person who is an author of items within Strathprints is uniquely defined. The application of these name authority files facilitates the retrieval of the complete set of any one author's research output via an author browse menu.

Additional content comprises policy documents (based on ROAR[4] templates) and statistical reports[5] reflecting the number of downloads per item. Support files including training documentation are also held within the system as html pages.

The Strathprints team has not attempted to capture the current information environment of the repository using a formal tool such as e.g. DAF (Digital Audit Framework) or DRAMBORA (Digital Repository Audit Method Based on Risk Assessment).

4.1 Stewardship requirements

As the central repository for the University of Strathclyde, Strathprints has a requirement to retain all research output (or metadata relating to such output) produced by University staff, and to and provide access to it where possible. The rationale behind institutional repositories is to make research and other output available on an open access basis. Copyright restrictions may apply to some items preventing them being openly available. In such cases, metadata records only will be made publicly available, with links to full-text items that are available elsewhere.

Although items held within Strathprints are not held due to legal requirements, they are retained to fulfil the requirements of the University's research policy. Associated metadata and administrative files are retained to assist with searching/browsing and management of the repository content.

[4] http://roar.EPrints.org/
[5] http://strathprints.strath.ac.uk/es/index.php?action=cumulative_usage;range=all

4.2 Management team structure

Strathprints is managed by a team set-up, with current staffing spread across the University Library and the Centre for Digital Library Research (CDLR).

Specific roles include:

- The Institutional Repository Manager: Alan Slevin,
- CDLR Team: Alan Dawson (technical support, software development, metadata importing and manipulation); Emma McCulloch (metadata development). CDLR also maintain the Strathprints server and a development server, which also serves as an emergency back-up server.
- Library Staff: Checking of, and approval of, metadata and full text as deposited in the submission buffer.

4.3 Workflow

As noted in the Introduction, the workflow to deposit a paper in Strathprints begins either with an individual researcher/academic or a departmental proxy. They complete a metadata record to describe the item, completing set fields as appropriate. Once completed, and the full-text paper uploaded if available, the record and any corresponding files are transferred to the repository's submission buffer. Qualified and trained library staff then assess each record, supplementing and editing the metadata record as they see fit and according to standards and conventions used. The copyright status of each item is checked (e.g. by consulting records within SHERPA/RoMEO, relating to restrictions, conditions and embargos), and a copyright cover sheet is added to each document. Library staff are able to approve items by moving them from the submission buffer to the live archive. An automated process of indexing then ensures the items are available for searching and browsing.

The Strathprints server is currently located within CDLR, though there are plans to move this to the central IT support service so that it is managed in line with other core university services.

5. Phase 2: Evaluating techniques to green IM

Three techniques of a list of seven presented were deemed relevant to Strathprints. These are ECM (Enterprise Content Management), de-duplication and version control.

5.1 ECM

ECM is likely to bring benefits to the management of Strathprints, although an ongoing University project - RIMS (Research Information Management System)[6] - is looking to bring together all research administration systems, possibly using a software package called PURE. It is possible that such a system may meet the functionality that would be afforded by ECM.

[6] http://www.strath.ac.uk/rims/

5.1.1 Local benefits

One benefit of ECM is that it would facilitate better integration of email correspondence and repository content. A large volume of email is currently retained within local email clients, as evidence of author permissions to archive and to retain a record of exchanges relating to specific publications where necessary. Management issues occasionally arise due to the disassociation between the publications themselves and email files relating to them. It would improve the robustness of the service (from a management perspective) if these two types of record could be retained within the same system. Management would be improved since less work would require to be replicated; on occasion, required information cannot be located efficiently within the email client (usually outlook) due to a limited search facility.

5.1.2 Local disadvantages

A potential difficulty lies within the required open access status of Strathprints, as it is unclear whether or not this status could be maintained with an ECM set-up. A second difficulty might be that library expertise, which ensures high quality metadata records are added and full text content is not bound by copyright restriction, becomes lost due to the centralised model. Experience with the RM database, as mentioned in the introduction section above, shows that administration of bibliographic information and determination of copyright status of individual items is better handled by qualified library staff, resulting in a more consistent, higher-quality records. The continued involvement of library staff therefore, would require to be built into the workflow of any centralised model.

5.1.3 Institutional benefits

Wider organisational benefits resulting from the integration of repository content and associated email files include centralisation of repository related material at institutional level, increased transparency and decreased need for interoperability across related systems. It is also likely that this approach would mean preservation would be addressed at an institutional level, since work into investigating how best to preserve resources within independent and local systems would be greatly reduced.

It is also thought likely that more effective/efficient compliance re FOI might be achieved as a result of the implementation of ECM, due to increased centralisation of an institution's information resources. Provided it is effectively managed within the ECM, identification of eligible information in response to an FOI request could potentially be helpful when compiling responses.

It is thought likely that the use of ECM will result in a reduction in digital storage capacity, through the improved ability to identify information being held. This improved visibility of information should facilitate the identification of duplicate files and low value information. Only the active deletion of duplicates and low value information will result in reductions in storage space however; the use of ECM could potentially provide a stepping stone toward this green outcome.

5.1.4 Institutional disadvantages

It is possible that stewardship requirements may be compromised however. Due to centralisation, there is an increased chance of potentially diluting responsibility, so close control over information

management would be required. Service Level Agreements (SLAs) would help to ensure effective stewardship throughout an institution.

5.2 De-duplication

Rationalisation of duplicate files within Strathprints is facilitated by the EPrints software itself. For example, drop down menus highlight the presence of e.g. duplicate conference papers and journal articles. Authors sometimes submit both a conference presentation and a journal article bearing the same title (and, obviously, author). EPrints will flag up such overlap via the drop down menu facility, which will recognise duplicate entries in the relevant metadata fields.

It would be wrong, however, to delete one such file, since both outputs may count towards an author's official research output. The two records, although bearing identical titles and other fields, do not represent the same item. This content should therefore be retained within the repository.

The presence of these apparent duplicates may be confusing for a user. A policy decision to favour a journal article over an identically titled conference presentation has been taken, since the journal article is considered to be an official finished and, usually, peer-reviewed, research output. Although both 'versions' will be retained within the repository, since they should be included in a return to the REF (Research Evaluation Framework) and on author's CVs, de-duplication of search results is set-up. That is, the system uses metadata to de-duplicate search results so that searches do not return identically titled items. A metadata field enabling items to make reference to each other is populated for these items.

De-duplication is applicable to the range of different item-types held within Strathprints. Different metadata fields are populated for different time types but the identification of duplicates currently involves title and URL fields only, two fields included in records for all types of output. Regular checks are run to identify and remove duplicate items from the system, with precise matching on external URLs and some fuzzy matching on item titles. If a duplicate item is still in the submission buffer then it is removed entirely from the system. However, if the duplicates have both been added to the live repository then one of them is marked as 'deleted' and is removed from public display, but both records are retained within the system unless extra steps are taken to delete one of them manually from the EPrints database.

Beyond the repository platform itself, duplication of files across the institution may be caused by authors retaining copies of their publications on their personal computers or on departmental servers or websites. Unless sufficiently robust methods of storage, access, disaster recovery and preservation is implemented, and associated policies promoted by senior management, it is unlikely that staff will be sufficiently trusting of the repository model; this may result in them being insufficiently confident to delete duplicate copies from their local machines. To overcome this, changes to working practices are required and an effective change management process introduced, together with the reassurance that the repository infrastructure is able to provide ready access to material, in a way as straightforward as it would be for a researcher to lift a publication from the desktop of his/her own local machine.

In addition to de-duplicating entire files within a repository, or across repositories, duplication of individual pieces of content within files may be minimised (intra-file de-duplication). It is possible,

for example, to remove common content from all resources within a repository and to store one centralised or master copy of that specific piece of content. Pointers would then be inserted to the master copy from all items incorporating that particular piece of content. At the point of access by a user, aggregation will take place to import common content into individual files as appropriate. This can be achieved within EPrints using an existing web service. Strathprints plans to implement this in future, development time permitting.

5.2.1 Local benefits

De-duplication serves to streamline the repository content. It also provides benefits when it comes to harvesting, by e.g. OAIster, since duplicate content will not be harvested and require to be de-duplicated by external parties, who may not have the full information required to determine which is the best (in terms of quality, accuracy, completeness etc) version to retain. Better that this sort of editorial 'house-keeping' is carried out locally, where authors within the institution can be consulted to provide clarification on their outputs if necessary.

The implementation of information stewardship requirements will be enhanced by promoting de-duplication. One definitive metadata record or full-text file per item will be retained, in line with copyright policy for the particular journal/conference in which it featured.

In terms of intra-file de-duplication, the current manual process of adding copyright coversheets to individual items will be made more efficient as a result of increased automation. Server space will also be reduced since duplicate content will not be held for each individual item; only one copy of such content need be retained.

5.2.2 Local disadvantages and institutional benefits

Distinction should be made here between de-duplication within a repository and de-duplication across different repositories. De-duplication within a repository will bring local and organisational benefits; the same practice across different systems within an institution or across different systems held in different institutions, may introduce limitations. In contrast to benefits from streamlining content, one school believes that duplication of records and/or items is positive since multiple records increase the visibility of research outputs. Recent discussion on the JISCmail Repositories discussion list[7] suggests that "Having duplicate copies is a good idea for preservation and also if people are likely to access one collection but not another", claiming that this supports the need for "both institutional and subject repositories". Evidence for this within Strathclyde, is illustrated by Strathprints' current lack of interoperability with other library systems. This means that duplication is deemed attractive across systems, to maximise users' chances of accessing relevant material. It follows that duplication between Strathprints and the library OPAC may prove beneficial. The JISC-funded OCRIS[8] project is looking at this issue and its outcomes will be of interest. Likewise, duplication between Strathprints and the Digitool repository may be favourable. Potential difficulties caused by such duplication can be minimised by a robust system of managing item IDs (unique identifiers) across different platforms, since this will make the identification of duplicates straightforward across systems.

---

[7] Jenny Delasalle, University of Warwick, 'On Duplicate copies', 1st October 2009, https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0910&L=JISC-REPOSITORIES&T=0&F=&S=&P=3027

[8] http://cdlr.strath.ac.uk/ocris/

Intra-file de-duplication will result in staff time savings and, hence, more quickly achieved institutional branding, particularly where repository content is harvested by external parties.

It is unlikely that compliance (FOI etc) will be facilitated through the use of de-duplication methods since, by its very nature, all content within an institutional repository is already located within the public domain. It may increase the rate at which eligible material can be identified and packaged however.

5.2.3 Institutional disadvantages

Additional concerns include the ability to provide a comprehensive list of all qualifying research output for REF purposes and the ability to generate staff CVs from the repository. Duplication would result in the generation of poor quality lists of publications, yet duplication of titles (e.g. for a related conference presentation and journal article) should be included in such lists.

A recent question arising in relation to the issue of duplication is how to handle the records of an author who has left the institution? Should they be retained? Should they be deleted to avoid two institutions submitting the same papers to the REF? The wider question here is: should the repository reflect research output as it occurred, retaining material for those in post within the institution at the time of publication, or should it reflect research output according to 'real time' circumstances?

5.3 Version control

Depending on the stage of deposit within the publication process and in order to assert compliance with publisher and/or funder copyright policies, different versions of the same item may appear within Strathprints. Preprints are accepted, which means that authors may upload a version of a paper at the same time as that paper is submitted to a journal for consideration for future publishing.

There is no systematic workflow in place to remove preprints when corresponding postprints become available. Ad hoc checks are undertaken to carry this out. One approach to establishing a more comprehensive means of 'weeding' preprints, would be to devolve this responsibility to proxies (trained departmental staff) but it is unclear whether or not this would be effective.

Different versions of an item can be linked by in-built EPrints functionality.

One difficulty in establishing policy relating to version control is that it is not always clear what constitutes a 'version', or exactly what constitutes a preprint, postprint, author final draft, and so on.

One long term development in the publishing arena that may be facilitated by the ability of EPrints repositories to accommodate different versions of an item is that it may be feasible to handle a proportion of the peer-review process within the repository platform. Submitted versions (preprints) could be uploaded and reviewers invited to undertake reviews within the repository itself. Reviewed and amended versions, and eventually published versions, could then be uploaded and made available.

Research into version control within repositories is ongoing. See, for example, JISC's Version Identification Framework[9].

It is considered unlikely that version control will result in reduced digital storage within an institution. The technique will promote better management of versions but there will not necessarily be any weeding out of different versions (unless they constitute duplicate files).

Software exists to improve the storage efficiency of different versions of a document. This is not particularly relevant to the repository scenario since items deposited, although possibly versions of the same final output, are final, recognised versions in themselves – preprint, postprint etc. The purpose of storing them is to provide a record of the publishing process for that particular piece of research, rather than storing changes between items only. Version control, within the present case study, will be considered in relation to managing different complete versions of a single publication.

5.3.1 Local benefits

By incorporating a workflow, or even plug-in software, to handle version control, different versions of the same item would be easily identified and linked. This would ensure the user is clearly guided through alternate versions and would also ensure that a final published version (or preprint) is accessed, where it exists, through 'signposting' from non-final versions.

One benefit of a repository holding different versions, particularly in EPrints, is that it may enable the provision of content otherwise legally restricted to being made openly accessible. For example, a journal policy may prevent authors from uploading publisher final versions to a repository. One work-around to this, to enable an author to increase the visibility of his/her research through making it widely accessible, is to upload a pre-print, which has no restriction imposed by the subsequent publisher of the works.

Information stewardship requirements should still be met, irrespective of the implementation of a policy on version control. It is probable that restrictions imposed by journal publishers and/or funders, however, will compromise the extent to which different versions may be made available.

5.3.2 Local disadvantages

Unless versions are carefully managed and relationships imposed between them using metadata, the most recent versions, and hence the best versions to cite, may not be easily identifiable.

5.3.3 Institutional benefits

The generation of staff CVs from Strathprints will be optimised, with only the final versions of items being listed where they exist. Metadata will ensure the most recent version of an item, as uploaded or described in line with copyright policy, is included within CVs.

Clearer information on REF activity can be gleaned from the repository where version control is in place.

[9] http://www.jisc.ac.uk/whatwedo/programmes/reppres/vif.aspx

5.3.4 Institutional disadvantages

Much discussion surrounds the citing of repository items, with many of the view that citation of a non-final version is somehow unrepresentative of the scholarly communication chain, since non peer-reviewed versions may be cited where more robust versions exist. One view is that this will compromise the profile of a researcher and, in turn, an institution. A contributor to the JISCmail Repositories discussion list[10], however, doesn't view this as an issue claiming that "most people will read and cite the final publisher's version anyway", whether or not this is the version that is held or described within a repository. They may first encounter an item within a repository but will use the definitive source of the material if including a citation within their own works. The contributor concludes that "there can be no harm to the author's profile in having numerous versions of their work on the web, so long as they all point to the final published one as well". It remains unclear, therefore, whether or not the debate over the citation of non-publisher versions will cause any long-term difficulties or have implications for the REF.

5.4 Inappropriate/infeasible GIM techniques

From the list of techniques identified by the GIM framework developed within the project, quotas are considered inappropriate/infeasible for Strathprints. University policy holds that the entire institutional research output must be made available on an open access basis. Quotas that would potentially limit the required storage capacity for upholding this policy cannot therefore be introduced.

6. Phase 3: Assessing costs and benefits

De-duplication was selected as the choice of GIM technique considered most appropriate for Strathprints. As described before, this technique may relate to duplication of items or de-duplication of content within items (intra-file de-duplication). The former is already being undertake in Strathprints and the latter is deemed to be a feasible technique to implement.

6.1 Working practices

Currently, the de-duplication of journal papers and conference presentations within Strathprints' search results is managed by a mixture of automated and manual methods. Each item is checked manually, following the automated matching of titles and URLs within the Strathprints database. Changes to current working practices could be introduced to try to avoid the upload of duplicate files at the point of entry. This would require to be undertaken by departmental proxies. It is not a straightforward issue however, since duplication of journal papers and conference papers for CV purposes is desirable, since each reflects a valid research output, but this is undesirable within the repository itself, since users tend to want access to the more 'formal' output – usually the journal article – which they would be more comfortable citing within their own works.

The planned intra-file de-duplication provides the opportunity to create more fluid workflows. Currently, library staff undertake repository work on a fairly ad-hoc basis, fitting it in around current

[10] Jenny Delasalle, University of Warwick, 'On Duplicate copies', 1st October 2009, https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0910&L=JISC-REPOSITORIES&T=0&F=&S=&P=3027

duties, which may vary greatly from day to day. The automated addition of coversheets, using master content where feasible, would reduce the workload on such staff. A program could extract the bibliographic record from the metadata record, and append this to the coversheet, removing the need to extract this on a manual basis. Shared content would be pointed to. What is not easily incorporated into this procedure is checking of the copyright status of actual content. If author final drafts are available, the wording on the coversheet should be fairly standard. This will not always be the case though, so it is likely that checks will have to be undertaken, on a manual basis, to avoid copyright restrictions being flouted.

The introduction of further de-duplication techniques is likely to introduce added responsibility for proxies and increased training requirements.

6.2 Business implications

It is thought likely that the introduction of further de-duplication techniques will increase the institutional profile of Strathclyde, facilitate the process involved in making a submission to the REF, improve means of legislative compliance, increase the consistency with which copyright declarations are implemented and result in a long-term reduction in costs as a result of decreased staff time required to manage related tasks. Within the institution, robust procedures will also serve to engender trust in the repository among the academic community, which should result in higher participation rates and overall increased success of Strathprints.

6.3 Digital footprint

De-duplication of files and of individual pieces of content will result in reduced storage space requirements. The introduction of an additional web service will have minimal impact on Strathprints' carbon footprint. A reduction in the number of machines used to administer de-duplication through the automated addition of coversheets and aggregation of common content is a likely outcome, as there will be no requirement to add pdf coversheets to every individual full-text item held within the repository. This is currently a large cost, since several staff members could be undertaking this task simultaneously across a number of machines. Reduced disk space will be achieved through the storage of a master copy of common content, rather than storing a copy of this for every individual item. This should make searching more efficient as fewer PDF files would mean that less indexing will be required.

6.4 Change management

An alternative workflow would be required to accommodate de-duplication of files at entry level, as handled by proxies. To accommodate automated de-duplication of common content, an initial saving of that content on a centralised server would be required; a change in workflow to automatically add coversheets to qualifying items (i.e. full text author final drafts) would also be needed.

Changes to training programmes would be required, although the nature of such changes are currently unknown.

Changes to advocacy are likely, since proxies would become responsible for promoting the use of the repository to, and ensuring contributions were made by, members of their department.

6.5 Evaluating costs/benefits

Benefits are likely to be found in relation to staff time (and hence financial savings), compliance, increased visibility of individual researchers, institutional promotion and carbon footprint as a result of more efficiently used disk space.

Cost savings could be evaluated by comparing the levels of work undertaken by relevant parties now and six months after the implementation of a de-duplication technique i.e. the web service. This would relate to the roles of both library staff and departmental proxies. The extent of change to workloads and workflows could be assessed on this basis. De-duplication may result in short-, medium- and long-term benefits. These could be identified and evaluated over appropriate time periods. Spot-checks to ascertain whether de-duplication is in fact improving the service should be run at regular intervals to assess the technique's success. This could be done systematically using the author or departmental browse lists. This would facilitate measures of whether storage space was being saved as a direct result of this technique.

Costs are likely to be found in relation to development time (customisation), technical support and training (short-term).

The level of customisation required to implement an automated process of adding copyright coversheets, which point to common and centralised content, is currently unknown. This would need to be undertaken before the associated costs and benefits could be accurately assessed.

The introduction of the web service facilitating de-duplication is likely to require increased technical support (ongoing), although since it will be integrated into the same server as the EPrints repository itself, this should not be significant. Infrastructure costs would remain the same. Increased training would be required on a short-term basis to familiarise the proxies with the new workflow.

There are likely to be competing costs and benefits for different parties within the institution. For example, departmental proxies may have additional costs in terms of workload, but the benefits this extra workload would bring to the institution would far outweigh this cost. For example, an accurate collection of research output and accurate metadata records would bring significant savings to an institution while compiling a response to the REF.

The perspectives of researchers may also influence the extent of costs and benefits, as they are considered in relation to the repository as a whole. For example, the difference between the prominence of researchers might influence their level of participation. Highly ranked departments (in RAE terms) may be keen to deposit their work in Strathprints to create improved visibility, and increasing the reputation of their department (and institution further). However, a lower ranking department may see contributing to Strathprints as a negative benefit. That is, they will be dissuaded from using the repository as a means of publicising lower RAE rankings. A threshold of users is required before Strathprints will be considered a total success.